Divergence Loss and the Stein Phenomenon

Malay Ghosh
University of Florida

June 19, 2019

Outline

- Introduction
- Squared Error Loss
- Divergence Loss
- Inadmissibility Results
- Stein's Harmonic Prior
- A Two Sample Problem
- Final Remarks

Introduction

- The natural estimator of the normal mean, the sample mean, is its UMVUE, the MLE, and also the best equivariant estimator under translations of the sample space.
- This estimator is minimax under a very general class of losses in any arbitrary dimension.
- Blyth (1951) proved the admissibility of the sample mean in the normal case for a very general class of losses, including but not limited to the squared error loss, in one dimension.
- Stein (1956) made the surprising discovery that although the sample mean is an admissible estimator of the normal mean in one or two dimensions under squared error loss, it is not so in three or higher dimensions under the same loss.

- Baranchik (1970), under squared error loss, provided a very general class of minimax estimators including the James-Stein estimator which dominates the sample mean in three or higher dimensions.
- A very important subclass of the Baranchik class of estimators, namely proper Bayes minimax estimators under certain hierarchical priors, was found by Strawderman (1971).
- Efron and Morris, in a series of articles, provided an empirical Bayes (EB) interpretation of the James-Stein estimator.
- Extending their ideas further, the entire Baranchik class of estimators, including those of Strawderman, can be given simple EB interpretation.

- James-Stein estimators or their variants which dominate the sample mean continue to do so when it comes to prediction of a future observation form the normal distribution with the same mean under squared error loss.
- This is a consequence of a more general result which establishes duality between estimation and prediction under squared error loss.
- Robert (2001) points out that often it is natural to use losses which compare directly the densities $f(\cdot|\boldsymbol{x})$ and $f(\cdot|\boldsymbol{a})$, where θ is the true parameter.
- Robert refers to such losses as "intrinsic losses".

- The two most well-used divergence measures between two densities are the Kullback-Leibler (KL) or the entropy distance and the Bhattacharyya-Hellinger (BH) distance (Hellinger, 1909; Bhattacharyya, 1943).
- However, if $\boldsymbol{X} = (X_1, \ldots, X_p)^T \sim \mathrm{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$, the KL distance between $f(\cdot|\boldsymbol{\theta})$ and $f(\cdot|\boldsymbol{a})$ is

$$E_{\boldsymbol{\theta}}\left[\log \frac{f(\boldsymbol{X}|\boldsymbol{\theta})}{f(\boldsymbol{X}|\boldsymbol{a})}\right] = \frac{1}{2v_x}||\boldsymbol{\theta} - \boldsymbol{a}||^2.$$

- Hence, point estimation based on the KL loss is tantamount to squared error loss for known v_x.

- KL and BH losses are special cases of a more general divergence loss considered in other contexts (Renyi, 1961; Amari, 1982; Cressie and Read, 1984).
- This loss is given by

$$L_\beta(\theta, \boldsymbol{a}) = \frac{1 - \int f^{1-\beta}(\boldsymbol{x}|\boldsymbol{\theta}) f^{\beta}(\boldsymbol{x}|\boldsymbol{a}) d\boldsymbol{x}}{\beta(1-\beta)}$$

- The above loss is to be interpreted as its limit when $\beta \to 0$ or $\beta \to 1$.
- $\lim_{\beta \to 0} L_\beta(\boldsymbol{\theta}, \boldsymbol{a}) = E_{\boldsymbol{\theta}}[\log \frac{f_{\boldsymbol{\theta}}(\boldsymbol{X})}{f_{\boldsymbol{a}}(\boldsymbol{X})}]$.
- $\lim_{\beta \to 1} L_\beta(\boldsymbol{\theta}, \boldsymbol{a}) = E_{\boldsymbol{a}}[\log \frac{f_{\boldsymbol{a}}(\boldsymbol{X})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})}]$.
-

$$\begin{aligned} L_{1/2}(\boldsymbol{\theta}, \boldsymbol{a}) &= 4[1 - \int f^{1/2}(\boldsymbol{x}|\boldsymbol{\theta}) f^{1/2}(\boldsymbol{x}|\boldsymbol{a}) d\boldsymbol{x}] \\ &= 2 \int [f^{1/2}(\boldsymbol{x}|\boldsymbol{\theta}) - f^{1/2}(\boldsymbol{x}|\boldsymbol{a})]^2] d\boldsymbol{x}. \end{aligned}$$

- Recently Komaki (2001) and George et. al (2006) have considered improved minimax predictive densities under the KL loss.
- They have developed various shrinkage versions of predictive densities which dominate under this loss he Bayes predictive density with the uniform prior for a future observation conditionally independent of the sampled observations.
- George et. al (2006) have also explored various iinteresting duality results between multivariate estimation and prediction in the normal problem under this loss.

- In contrast, Ghosh, Mergel and Datta (2006) considered the general divergence loss, established duality between estimation and prediction, and showed that the general Baranchik class of estimators continues to dominate the sample mean under this general divergence loss.
- Moreover,the EB interpretation of the Baranchik class of estimators continues to hold under this loss.
- The key feature of this talk is to explore this duality between estimation and prediction under the general divergence loss.

Squared Error Loss

- Let $X_1, \cdots, X_p | \theta_1, \ldots, \theta_p$ be independent $\mathrm{N}(\theta_i, v_x)$, $i = 1, \ldots, p$, $v_x > 0$ known.
- We write $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.
- For $p \geq 3$, under the loss $L(\boldsymbol{\theta}, \boldsymbol{a}) = ||\boldsymbol{a} - \boldsymbol{\theta}||^2$,
  $\delta^{\mathrm{JS}}(\boldsymbol{X}) = [1 - \frac{(p-2)v_x}{||\boldsymbol{X}||^2}]\boldsymbol{X}$
  dominates $\boldsymbol{X}$.
- Let $S = ||\boldsymbol{X}||^2 / v_x$. Baranchik (1970) showed that the general class of estimators $\delta^{BA}(\boldsymbol{X}) = [1 - \frac{\tau(S)}{S}]X$ dominates $\boldsymbol{X}$ if (i) $\tau(S)$ is non-decreasing in S and (ii) $0 < \tau(S) < 2(p-2)$.
- If $\tau(S) = c$, $0 < c < 2(p-2)$, the optimal $c = p - 2$, which gives the James-Stein estimator.

- The general Baranchik class of estimators can be given an EB interpretation.
- $\boldsymbol{X}|\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$ and $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{0}, A v_x \boldsymbol{I}_p)$, v_x known, A unknown.
- The Bayes estimator of $\boldsymbol{\theta}$ under squared error loss is $(1-B)\boldsymbol{X}$, where $B = (1+A)^{-1}$.
- To estimate B, note that marginally $\boldsymbol{X} \sim N(\boldsymbol{0}, B^{-1} v_x \boldsymbol{I}_p)$.
- $S = ||\boldsymbol{X}||^2/v_x$ is complete sufficient for B.
- A general class of EB estimators of B is of the form $(1-\hat{B}(S))\boldsymbol{X}$.
- In particular, since marginally, $S \sim B^{-1}\chi^2_p$, the UMVUE of B is $(p-2)/S$.
- The resulting EB estimator of B is $(p-2)/S$ leading to the James-Stein estimator $(1-\frac{p-2}{S})\boldsymbol{X}$.

- An important class of Bayes estimators dominating $\boldsymbol{X}$ was provided by Stein (1981).
- Consider a general prior $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$.
- Let $m_\pi(\boldsymbol{X}) = \int \phi(\boldsymbol{X}|(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ be finite for all $\boldsymbol{X}$.
- Then the Bayes estimator of $\boldsymbol{\theta}$ is $\delta_\pi(\boldsymbol{X}) = \boldsymbol{X} + v_x \bigtriangledown \log m_\pi(\boldsymbol{X})$, where $\bigtriangledown$ denotes the gradient vector.
- Stein showed that if either m_π or $m_\pi^{1/2}$ is superharmonic, i.e. it has a negative Laplacian, then $\delta_\pi(\boldsymbol{X})$ dominates $\boldsymbol{X}$.
- In particular, Stein proposed the harmonic prior $\pi(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||^{-(p-2)}$.

- Duality between dominance under squared error loss over the estimator $\boldsymbol{X}$ of $\boldsymbol{\theta}$ and the predictor $\boldsymbol{X}$ of $\boldsymbol{Y}$ where $\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{\theta}$ and $E(\boldsymbol{X}) = E(\boldsymbol{Y}) = \boldsymbol{\theta}$.
- Suppose an estimator $e(\boldsymbol{X})$ of $\boldsymbol{\theta}$ dominates $\boldsymbol{X}$ under squared error loss. Then
- $$\begin{aligned} & E_{\boldsymbol{\theta}}||e(\boldsymbol{X}) - \boldsymbol{Y}||^2 \\ &= E_{\boldsymbol{\theta}}||\{e(\boldsymbol{X}) - \boldsymbol{\theta}\} - \{\boldsymbol{Y} - \boldsymbol{\theta}\}||^2 \\ &= E_{\boldsymbol{\theta}}||e(\boldsymbol{X}) - \boldsymbol{\theta}||^2 + E_{\boldsymbol{\theta}}||\boldsymbol{Y} - \boldsymbol{\theta}||^2 \\ &< E_{\boldsymbol{\theta}}||\boldsymbol{X} - \boldsymbol{\theta}||^2 + E_{\boldsymbol{\theta}}||\boldsymbol{Y} - \boldsymbol{\theta}||^2 \\ &= E_{\boldsymbol{\theta}}||\boldsymbol{X} - \boldsymbol{Y}||^2. \end{aligned}$$
- Conversely, suppose $E_{\boldsymbol{\theta}}||e(\boldsymbol{X}) - \boldsymbol{Y}||^2 < E_{\boldsymbol{\theta}}||\boldsymbol{X} - \boldsymbol{Y}||^2$.
- Then subtracting $E_{\boldsymbol{\theta}}||\boldsymbol{Y} - \boldsymbol{\theta}||^2$ from both sides, $E_{\boldsymbol{\theta}}||e(\boldsymbol{X}) - \boldsymbol{\theta}||^2 < E_{\boldsymbol{\theta}}||\boldsymbol{X} - \boldsymbol{\theta}||^2$.

Divergence Loss

- Let $\boldsymbol{X} \perp \boldsymbol{Y}|\theta$ with corresponding pdf's $p(\boldsymbol{x}|\theta)$ and $p(\boldsymbol{y}|\theta)$.
- Recall the divergence loss $L_\beta(\boldsymbol{\theta}, \boldsymbol{a}) = \frac{1-\int f^{1-\beta}(\boldsymbol{x}|\boldsymbol{\theta}) f^{\beta}(\boldsymbol{x}|\boldsymbol{a}) d\boldsymbol{x}}{\beta(1-\beta)}$.
- Let $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ denote the posterior pdf of $\boldsymbol{\theta}$ given $\boldsymbol{x}$.
- Lemma 1. Under the divergence loss and the prior π, the Bayes predictive density of $\boldsymbol{Y}$ is $\pi_D(\boldsymbol{y}|\boldsymbol{x}) = k^{1/(1-\beta)}(\boldsymbol{y}, \boldsymbol{x}) / \int k^{1/(1-\beta)}(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{y}$, where $k(\boldsymbol{y}, \boldsymbol{x}) = \int p^{1-\beta}(\boldsymbol{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}$.
- In the special case when $\beta \to 0$, namely the KL loss, $\pi_{KL}(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}$, the usual predictive pdf (Atchison, 1975).

- A useful result.
- Suppose $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are two independent p-component random vectors with pdf's f_1 and f_2 and distribution functions F_1 and F_2, respectively.
- Let h and H denote respectively the pdf and distribution function of $\boldsymbol{Z}_1 - \boldsymbol{Z}_2$.
- Then $h(\mathbf{0}) = \int f_1(\boldsymbol{z}) f_2(\boldsymbol{z}) d\boldsymbol{z}$.
- The proof follows by first oberving $H(\boldsymbol{x}) = \int F_1(\boldsymbol{x} + \boldsymbol{z}) f_2(\boldsymbol{z}) d\boldsymbol{z}$.
- Differentiating both sides with respect to $\boldsymbol{x}$, one gets $h(\boldsymbol{x}) = \int f_1(\boldsymbol{x} + \boldsymbol{z}) f_2(\boldsymbol{z}) d\boldsymbol{z}$, which leads to the value of $h(\mathbf{0})$.
- In particular, $\int \phi(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \phi(\boldsymbol{z}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) d\boldsymbol{z} = \phi(\mathbf{0}|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.
- Also, $L_\beta(\boldsymbol{\theta}, \boldsymbol{a}) = \exp[-\frac{\beta(1-\beta)}{2v_x} ||\boldsymbol{\theta} - \boldsymbol{a}||^2]$.

- Consider once again $\boldsymbol{Y} \perp \boldsymbol{X} | \boldsymbol{\theta}$ with $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\theta}, v_y \boldsymbol{I}_p)$ and $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$, $v_x (> 0)$ and $v_y (> 0)$ are known.
- $\boldsymbol{\theta}$ has the $\mathrm{N}(\boldsymbol{\mu}, A\boldsymbol{I}_p)$ prior. $(1 - B)\boldsymbol{X} + B\boldsymbol{\mu}$, where $B = v_x/(v_x + A)$.
- The goal here is to maximize $\int \exp\Big\{ - \frac{\beta(1-\beta)}{2v_x} ||\boldsymbol{\theta} - \boldsymbol{a}||^2 \Big\} \phi(\boldsymbol{\theta} | (1 - B)\boldsymbol{X} + B\boldsymbol{\mu}, v_x(1 - B)\boldsymbol{I}_p) d\boldsymbol{\theta}$ with respect to $\boldsymbol{a}$.
- This integral is proportional to $\exp\Big[- \frac{||\boldsymbol{a} - (1-B)\boldsymbol{X} - B\boldsymbol{\mu}||^2}{2v_x\{\beta^{-1}(1-\beta)^{-1} + 1 - B\}} \Big]$.
- This is maximized when $\boldsymbol{a} = (1 - B)\boldsymbol{X} + B\boldsymbol{\mu}$.
- Thus the Bayes estimator of $\boldsymbol{\theta}$ is $(1 - B)\boldsymbol{X} + B\boldsymbol{\mu}$, where $B = v_x/(v_x + A)$, same as squared error loss.

- For prediction, writing $(1-B)\boldsymbol{X}+B\boldsymbol{\mu}=\hat{\boldsymbol{\theta}}_B(\boldsymbol{X})$,

$$
\begin{aligned}
[\hat{p}_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{X})]^{1-\beta} &\propto [\int \phi^{1-\beta}(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p)\phi(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_B(\boldsymbol{X}), v_x(1-B)\boldsymbol{I}_p)d\boldsymbol{\theta}] \\
&\propto \ [\phi(\boldsymbol{y}|(1-B)\boldsymbol{X}+B\boldsymbol{\mu}, \{v_x(1-B)(1-\beta)+v_y\}\boldsymbol{I}_p]^{1-\beta}.
\end{aligned}
$$

- If $\pi(\boldsymbol{\theta})=1$, $B\to 0$.
- Then the Bayes estimator of $\boldsymbol{\theta}$ is $\boldsymbol{X}$.
- The Bayes predictive density of $\boldsymbol{Y}$ is $\phi(\boldsymbol{y}|\boldsymbol{X},(v_x(1-\beta)+v_y)\boldsymbol{I}_p)$.
- We denote by $\boldsymbol{\delta}_0$ the plug-in predictive density $\hat{p}_{pl}(\boldsymbol{y}|\boldsymbol{X})=\phi(\boldsymbol{y}|\boldsymbol{X}, v_y\boldsymbol{I}_p)$.

- For estimating the normal mean, the Bayes estimator of $\boldsymbol{\theta}$ under the prior $\pi(\boldsymbol{\theta}) = 1$ is $\boldsymbol{X}$ for the general divergence loss as well as squared error loss.
- Similarly, under the divergence loss, the Bayes predictive density of $\phi(\boldsymbol{y}|\boldsymbol{\theta}, v_y \boldsymbol{I}_p)$ under the prior $\pi(\boldsymbol{\theta}) = 1$ is $\hat{p}_U(\boldsymbol{y}|\boldsymbol{X}) = \phi(\boldsymbol{y}|\boldsymbol{X}, (v_x + v_y)\boldsymbol{I}_p)$.
- $\phi(\boldsymbol{y}|\boldsymbol{X}, (v_x + v_y)\boldsymbol{I}_p)$ has smaller risk than that of the plug-in predictor $\phi(\boldsymbol{y}|\boldsymbol{X}, v_y \boldsymbol{I}_p)$.
- $\boldsymbol{X}$ is a minimax estimator of θ under the general divergence loss in any arbitrary dimension.
- Similarly, $\hat{p}_U(\boldsymbol{y}|\boldsymbol{X}) = \phi(\boldsymbol{y}|\boldsymbol{X}, (v_x + v_y)\boldsymbol{I}_p)$ is a minimax predictor of $\phi(\boldsymbol{y}|\boldsymbol{\theta}, v_y \boldsymbol{I}_p)$ under the general divergence loss in any arbitrary dimension.
- While Stein type shrinkage estimators are targeted to dominate $\boldsymbol{X}$, the goal is to improve on $\hat{p}_U(\boldsymbol{y}|\boldsymbol{X})$ via shrinkage.
- This is what one wants to accomplish under the general divergence loss.

Inadmissibility Results

- $\hat{p}_U(\boldsymbol{y}|\boldsymbol{X})$ is an admissible predictor of $\phi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{I}_p)$ for $p = 1$.
- Its admissibility is an open question for $p = 2$. (?)
- $\hat{p}_U(\boldsymbol{y}|\boldsymbol{X})$ is an inadmissible predictor of $p(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p)$ for $p \geq 3$.
- The duality between estimation and prediction can be seen via the Baranchik class of estimators and predictors.
- Let $S = \|\boldsymbol{X}\|^2/v_x$.
- The Baranchik class of estimators for $\boldsymbol{\theta}$ is given by $\boldsymbol{\delta}^\tau(\boldsymbol{X}) = \{1 - \tau(S)/S\}\boldsymbol{X}$.
- Analogously, the Baranchik class of predictors for $\phi(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p)$ is given by $\phi(\boldsymbol{y}|\boldsymbol{\delta}^\tau(\boldsymbol{X}), ((1-\beta)v_x + v_y)\boldsymbol{I}_p)$.

- For the estimation problem, $L(\boldsymbol{\theta}, \boldsymbol{\delta}^{\tau}(\boldsymbol{X})) = \frac{1}{\beta(1-\beta)}\left[1 - \exp\left\{ - \frac{\beta(1-\beta)}{2v_x}\|\boldsymbol{\delta}^{\tau}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2\right\}\right]$.
- For the prediction problem,

$$L\left(\phi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{I}_p), \phi\left(\boldsymbol{y}|\boldsymbol{\delta}^{\tau}(\boldsymbol{X}), \{(1-\beta)v_x + v_y)\boldsymbol{I}_p\}\right)\right.$$

$$= \frac{1}{\beta(1-\beta)}\left[1 - \frac{(v_y)^{p\beta/2}\{(1-\beta)v_x + v_y\}^{p(1-\beta)/2}}{\{(1-\beta)^2 v_x + v_y\}^{p/2}} \exp\{- \frac{\beta(1-\beta)\|\boldsymbol{\delta}^{\tau}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2}{2\{(1-\beta)^2 v_x + v_y\}}\right].$$

- Theorem. Let $p \geq 3$. Assume (i) $0 < \tau(t) < 2(p-2)$ and (ii) $\tau(t)$ is a differentiable nondecreasing function of t. Then
(a) $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\tau}(\boldsymbol{X})) < R(\boldsymbol{\theta}, \boldsymbol{X})$ and
(b) $R\big(\phi(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p), \phi\left(\boldsymbol{y}\,|\boldsymbol{\delta}^{\tau}(\boldsymbol{X}), ((1-\beta)v_x + v_y)\boldsymbol{I}_p\right)\big)$

$< R\big(\phi\left(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p\right), \phi\left(\boldsymbol{y}|\boldsymbol{X}, ((1-\beta)v_x + v_y)\boldsymbol{I}_p\right)\big)$.

- The above result can immediately be extended to shrinkage towards an arbitrary regression surface.
- Suppose now that $\boldsymbol{X}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$.
- $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{K}\beta, A\boldsymbol{I}_p)$ where $\boldsymbol{K}$ is a known $p \times r$ matrix of rank $r(< p)$ and β is $r \times 1$ regression coefficient.
- Define $\boldsymbol{P} = \boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K})^{-1}\boldsymbol{K}^T$ and $R^* = ||\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X}||^2/v_x$.
- Let $p \geq r + 3$. ssume (i) $0 < \tau(t) < 2(p - 2)$ and (ii) $\tau(t)$ is a differentiable nondecreasing function of t. Then $\boldsymbol{X} - \{\tau(R^*)/R^*\}(\boldsymbol{X} - \boldsymbol{P}\,\boldsymbol{X})$ dominates $\boldsymbol{X}$ under the general divergence loss.
- A similar dominance result holds for prediction of $\phi(\boldsymbol{y}|\boldsymbol{\theta}, v_y\boldsymbol{I}_p)$.

- Consider now the case $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$, where both $\boldsymbol{\theta}$ and v_x are unknown.
- $S \sim \frac{v_x}{m+2}\chi^2_m$ independent of $\boldsymbol{X}$.
- This is the typical balanced one way ANOVA model where $\boldsymbol{X}$ is the vector of treatment means and S is a multiple of the error mean square.
- The goal continues to be estimation of $\boldsymbol{\theta}$ under the general divergence loss $L(\boldsymbol{\theta}, \boldsymbol{a}) = \frac{1-\exp\left[-\frac{\beta(1-\beta)}{2v_x}||\boldsymbol{\theta}-\boldsymbol{a}||^2\right]}{\beta(1-\beta)}$, $0 < \beta < 1$.
- Noting that $||\boldsymbol{X} - \boldsymbol{\theta}||^2/v_x \sim \chi^2_p$, the risk $R(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{1-[1+\beta(1-\beta)]^{-p/2}}{\beta(1-\beta)}$.

- Following Baranchik (1970) and Efron and Morris (1976), we consider the rival set of estimators $\boldsymbol{\delta}(\boldsymbol{X}) = \left(1 - \frac{\tau(||\boldsymbol{X}||^2/S)}{||X||^2/S}\right)\boldsymbol{X}$.
- Consider the loss $L(\boldsymbol{\theta}, \boldsymbol{\delta}(\boldsymbol{X})) = \frac{1 - \exp\left[-\frac{\beta(1-\beta)}{2v_x}||\boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\theta}||^2\right]}{\beta(1-\beta)}$.
- Theorem. Let $p \geq 3$. Assume
  (i) $0 < \tau(t) < 2(p-2)$ for all $t > 0$;
  (ii) $\tau(t)$ is a differentiable nondecreasing function of t for $t > 0$.
  Then $R(\boldsymbol{\theta}, \boldsymbol{\delta}(\boldsymbol{X})) < R(\boldsymbol{\theta}, \boldsymbol{X})$ for all $\boldsymbol{\theta} \in R^p$.

- Finally, let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ $(n \geq 2)$ be iid $\mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where both $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are unknown.
- Let $\boldsymbol{X} = \bar{\boldsymbol{Z}}$ and $\boldsymbol{S} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}})(\boldsymbol{Z}_i - \bar{\boldsymbol{Z}})^T$.
- The divergence loss is now given by

$$
\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{a}) &= \frac{1 - \int \phi^{1-\beta}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\Sigma}/n)\phi^{\beta}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{\Sigma}/n)d\boldsymbol{x}}{\beta(1-\beta)} \\
&= \frac{1 - \exp\left[-\frac{n\beta(1-\beta)}{2}(\boldsymbol{a}-\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\boldsymbol{\theta})\right]}{\beta(1-\beta)}.
\end{aligned}
$$

- Since $n(\boldsymbol{X} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\theta}) \sim \chi_p^2$, $R(\theta, \boldsymbol{X}) = \frac{[1-\{1+\beta(1-\beta)\}^{-p/2}]}{\beta(1-\beta)}$.

- Consider now the general class of estimators $\delta(\boldsymbol{X}, \boldsymbol{S}) = \left[1 - \frac{\tau(n\boldsymbol{X}^T\boldsymbol{S}^{-1}\boldsymbol{X})}{n\boldsymbol{X}^T\boldsymbol{S}^{-1}\boldsymbol{X}}\right]\boldsymbol{X}$.
- Consider the loss $L(\boldsymbol{\theta}, \delta(\boldsymbol{X}, \boldsymbol{S})) = \frac{1-\exp\left[-\frac{n\beta(1-\beta)}{2}(\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{S})-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{S})-\boldsymbol{\theta})\right]}{\beta(1-\beta)}$.
- Theorem. Let $p \geq 3$. Assume
  (i) $0 < \tau(t) < 2(p-2)(n-1)/(n-p+2)$;
  (ii) $\tau(t)$ is non-decreasing differentiable function of t.
  Then $R(\boldsymbol{\theta}, \delta(\boldsymbol{X}, \boldsymbol{S})) < R(\boldsymbol{\theta}, \boldsymbol{X})$.

Stein's Harmonic Prior

- Stein's harmonic prior yielded Bayes estimators of θ dominating $\boldsymbol{X}$ when $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{I}_p)$.
- To my knowledge, it was Komaki (2001) who first developed Stein type predictors for the density of a future observation under the normal set up.
- He showed in particular the dominance of the predictive density developed under Stein's (1981) harmonic prior over the classical best equivariant predictive density estimator as developed by Atchison (1975) under the Kullback-Leibler (KL) loss.
- George, Liang and Xu (2006) followed this work and exhibited this dominance under general superharmonic priors.
- They also demonstrated a very nice duality between estimation and prediction under superharmonic priors.

- Some details of the work of George et al. (2006).
- Let $\boldsymbol{Y} \perp \boldsymbol{X}|\boldsymbol{\theta}$ where $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\theta}, v_y \boldsymbol{I}_p)$ and $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, v_x \boldsymbol{I}_p)$.
- Let $\hat{p}_\pi(\boldsymbol{y}|\boldsymbol{X}))$ denote the posterior predictive pdf of $\boldsymbol{Y}$ under an arbitrary prior π.
- Let $m_\pi(\cdot)$ denote a generic notation for marginals.
- Let $W = (v_x^{-1}\boldsymbol{X} + v_y^{-1}\boldsymbol{Y})/(v_x^{-1} + v_y^{-1})$.
- Let $\hat{p}_u(\boldsymbol{y}|\boldsymbol{x})$ is the posterior predictive pdf of $\boldsymbol{Y}$ given $\boldsymbol{X}$ under the uniform prior $\pi(\boldsymbol{\theta})$; $v_w = (v_x^{-1} + v_y^{-1})^{-1}$.
- Result 1 (The basic identity)
  $\hat{p}_\pi(\boldsymbol{y}|\boldsymbol{x}) = [m_\pi(\boldsymbol{w}; v_w)/m_\pi(\boldsymbol{x}; v_x)]\hat{p}_u(\boldsymbol{y}|\boldsymbol{x})$
- Result 2. If the prior $\pi(\boldsymbol{\theta})$ is superharmonic, then m_π is superharmonic) or equivalently $m_\pi^{1/2}$ is superharmonic.

- Result 3.
  $R_{KL}(\boldsymbol{\theta}, \hat{p}_u(\boldsymbol{Y}|\boldsymbol{X})) - R_{KL}(\boldsymbol{\theta}, \hat{p}_\pi(\boldsymbol{Y}|\boldsymbol{X})) =$
  $\int_{v_w}^{v_x} \frac{1}{2v^2} \left\{ E_{\boldsymbol{\theta},v}[||\boldsymbol{X} - \boldsymbol{\theta}||^2 - ||\delta_\pi(\boldsymbol{X}) - \boldsymbol{\theta}||^2 \right\} dv,$
  where $\delta_\pi(\boldsymbol{X})$ is the Bayes estimator of $\boldsymbol{\theta}$ under the prior $\pi(\boldsymbol{\theta})$ relative to the quadratic loss.
- Thus the dominance property of the Bayes point estimator of $\boldsymbol{\theta}$ under squared error loss is inherited by the Bayes predictive densities under KL loss.
- Stein (1981) showed that if $\pi(\boldsymbol{\theta})$ is superharmonic, then $\delta_\pi(\boldsymbol{X})$ dominates $\boldsymbol{X}$ under squared error loss.
- By Result 3, the same dominance holds under KL loss.

A Two Sample Problem

- Simultaneous estimation of two population means when one suspects the two means to be nearly equal.
- Rather than estimating the two means separately, it may be more profoitable to shrink each mean towards a pooled mean.
- An EB approach handles this in an adaptive way rather than a preliminary test approach.
- Ghosh and Sinha (1988) proposed hierarchical empirical Bayes estimators dominating individual sample means under squared error loss.
- The same phenomenon holds under a general divergence loss.

- $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and S are mutually independent random variables with $\boldsymbol{X}_i \sim \mathrm{N}_p(\boldsymbol{\mu}_i, \sigma^2 \boldsymbol{V}_i)$, $i = 1, 2$; $S/\sigma^2 \sim \chi^2_n$.
- Notation: $f(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2)$ joint pdf of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ under the above model.
- Divergence loss : $L_\beta(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{a}_1, \boldsymbol{a}_2) = \frac{1 - \int f^{1-\beta}(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2) f^\beta(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{a}_1, \boldsymbol{a}_2, \sigma^2) d\boldsymbol{x}_1 \boldsymbol{x}_2}{\beta(1-\beta)}$.
- g priors: $\boldsymbol{\mu}_i \sim \mathrm{N}_p(\boldsymbol{\nu}, \tau^2 \boldsymbol{V}_i)$, $i = 1.2$ independently
- $\boldsymbol{\nu} \sim$ uniform(R^p)
- The HB estimators of the $\boldsymbol{\mu}_i$ are given by $\hat{\boldsymbol{\mu}}_i^{HB} = \boldsymbol{X}_i - B(\boldsymbol{x}_i - \hat{\boldsymbol{\nu}})$, where $B = \sigma^2/(\sigma^2 + \tau^2)$ and $\hat{\boldsymbol{\nu}} = (\boldsymbol{V}_1^{-1} + \boldsymbol{V}_2^{-1})^{-1}(\boldsymbol{V}_1^{-1}\boldsymbol{Y}_1 + \boldsymbol{V}_2^{-1}\boldsymbol{Y}_2)$.

- First estimate B for known σ^2.
- Marginally $\boldsymbol{X}_1 - \boldsymbol{X}_2 \sim \mathrm{N}_p(\mathbf{0}, (\sigma^2/B)(\boldsymbol{V}_1 + \boldsymbol{V}_2)$.
- $(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T(\boldsymbol{V}_1 + \boldsymbol{V}_2)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)/\sigma^2 \sim B^{-1}\chi_p^2$.
- $E(1/\chi_p^2) = (p-2)^{-1}$ for $p \geq 3$.
- $\hat{B} = \frac{(p-2)\sigma^2}{(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T(\boldsymbol{V}_1 + \boldsymbol{V}_2)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)}$.
- Estimate σ^2 by $\hat{\sigma}^2 = S/(n+2)$.
- $F = (\boldsymbol{X}_1 - \boldsymbol{X}_2)^T(\boldsymbol{V}_1 + \boldsymbol{V}_2)^{-1}(\boldsymbol{X}_1 - \boldsymbol{X}_2)/\hat{\sigma}^2$.

- Consider the shrinkage estimator $\hat{\boldsymbol{\mu}}_i(\phi) = \boldsymbol{X}_i - \frac{\phi(F)}{F}(\boldsymbol{X}_i - \hat{\boldsymbol{\nu}})$, $i = 1, 2$.
- Theorem. $(\hat{\boldsymbol{\mu}}_1(\phi), \hat{\boldsymbol{\mu}}_2(\phi))$ dominates $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ under the general divergence loss for $p \geq 3$ if
  (a) $0 < \phi(t) < 2(p-2)$ for all $t > 0$;
  (b) $\phi(t)$ is a differentiable nondecreasing function of t for $t > 0$.
- The basic technique for proving this result is to introduce $\boldsymbol{W} = (\boldsymbol{V}_1^{-1} + \boldsymbol{V}_2^{-1})^{-1}(\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{V}_2^{-1}\boldsymbol{X}_2)$ and exploit the independence of $\boldsymbol{W}$ with $\boldsymbol{Z} = \boldsymbol{X}_1 - \boldsymbol{X}_2$.

Final Remarks

- Stein's 1956 paper was a major breakthrough in statistics, and its impact is felt even today.
- Stein's original results and many of their variants are continuously used both in the theory and application of statistics.
- The Stein phenomenon for prediction is of more recent origin.
- The present talk has focused on the Stein phenomenon for both estimation and prediction, and has pointed out the duality between the two in several contexts.